CH5350: Applied Time-Series Analysis

Maximum Likelihood and Bayesian Estimators

Arun K. Tangirala

Department of Chemical Engineering, IIT Madras

Introduction

We now turn to a fundamentally different approach to the estimation problem. Unlike the LS method, this estimator assumes the data to be generated by a probabilistic mechanism, possibly combined with some dynamics.

The key idea is to reverse "guess" the probability function that is "most likely" to have generated the given observations with or without prior knowledge of the parameters. In the former case, we are led to **maximum likelihood estimators** while in the latter we encounter **Bayesian estimation** methods.

Likelihood

The concept of likelihood is closely related to that of *conditional probability* (Bayesian ideas), but is *not* identical to it.

For any probabilistic event associated with a RV Y, we can compute the probability $\Pr(y_1 < Y < y_2)$ given its probability density function $f(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter vector characterizing the density function. Now the situation is that we are given a set of observations $\mathbf{y}_N = \{y[0], y[1], \cdots, y[N-1]\}$ and we are interested in "guessing" or estimating (i) the *form* of the density function (*model*) and (ii) the parameters of that function.

Obviously there exist infinite possibilities for $\boldsymbol{\theta}$ for a chosen $f(\mathbf{y}|\boldsymbol{\theta})$. A criterion is required to select the optimal parameters.

Maximum likelihood approach

Fisher, (1922) proposed that the optimal parameter is that one which maximizes the probability of occurrence of $\mathbf{y}$ in the range $\mathbf{y}_N$ to $\mathbf{y}_N + d\mathbf{y}$. **We are not maximizing probability here, but rather asking which p.d.f. (or parameter vector) would have made $\mathbf{y}_N$ most likely.**

From these ideas takes birth the concept of *likelihood* function $l(\boldsymbol{\theta}|\mathbf{y}_N)$. Essentially, given the observations $\mathbf{y}_N$, the p.d.f. $f(\mathbf{y}_N; \boldsymbol{\theta})$ is simply a function of the parameters $\boldsymbol{\theta}$. The likelihood function that we wish to maximize is proportional to the p.d.f. since the probability of the observation set taking values within an interval $d\mathbf{y}$ is proportional to the p.d.f. That is to say,

$$l(\boldsymbol{\theta}|\mathbf{y}_N) \propto f(\mathbf{y}|\boldsymbol{\theta}) \tag{1}$$

Maximum likelihood estimation

For computational reasons, we make two choices:

1. Neglect the proportionality constant in (1), so that

$$l(\boldsymbol{\theta}|\mathbf{y}_N) = f(\mathbf{y}|\boldsymbol{\theta}) \tag{2}$$

2. Maximize the log-likelihood $L(\boldsymbol{\theta}|\mathbf{y}_N) = \ln l(\boldsymbol{\theta}|\mathbf{y}_N)$ instead of the likelihood function, because it makes the problem numerically tractable.

Then, the ML estimator of $\boldsymbol{\theta}$ given observations $\mathbf{Z}_N = \{\mathbf{y}_N \cup \mathbf{u}_N\}$.

$$\boxed{\hat{\boldsymbol{\theta}}_{ML} = \arg \min_{\boldsymbol{\theta}} \text{-} \ln l(\boldsymbol{\theta}|\mathbf{y}_N) \qquad \text{or} \qquad \hat{\boldsymbol{\theta}}_{ML} = \arg \min_{\boldsymbol{\theta}} \text{-} L(\boldsymbol{\theta}|\mathbf{Z}_N)} \tag{3}$$

MLE procedure

A three-step procedure for the formulation and solution to a general ML estimation from N observations of a process $y[k]$ is described next.

It is useful to imagine the observations to be made up of a stochastic term $v[k]$ and/or a deterministic component $x[k]$.

$$y[k] = x[k] + v[k] \tag{4}$$

- In certain texts, the deterministic component is also said to be the conditional mean since $E(y[k]) = x[k]$ whenever $v[k]$ is zero-mean.
- In TSA, $x[k] = 0$ (or constant) for stationary signals.

MLE procedure

1. *Assume a density function*: Assume a suitable density function $f(\mathbf{v})$ for the stochastic component (typically a Gaussian).
2. *Construct the likelihood function*: Postulate a model (mostly dynamic) for the deterministic component (wherever applicable). Putting together the models for $x[k]$ and $v[k]$, construct the density function of $\mathbf{y}$ and hence the likelihood for $\boldsymbol{\theta}$.
3. *Solve the optimization problem*: Set up the optimization problem with any additional constraints that may be have to be placed. Solve it using a suitable algorithm (typically a numerical solver).

Remarks

- In order to carry out the last step, *i.e.*, to determine the optimum, **the density function should satisfy the regularity conditions**.
- When the p.d.f. is not regular, the derivative of the likelihood does not exist and the optimum has to be determined by inspection. An example of this case is the estimation of parameters of a uniform density function.
- Further, since MLE is a non-linear optimization problems, multiple solutions may exist. It is then necessary to verify that the solution indeed corresponds to the minimum by evaluating the second-derivative.
- Finally, as with any estimation technique, compute the errors in the resulting estimates.

Example 1: MLE of mean and variance

Mean and Variance estimation of a GWN process

Given N observations of a constant signal corrupted with noise,

$$y[k] = c + e[k] \tag{5}$$

estimate the the steady-state value c and the variance of the measurement noise σ_e^2.

Example . . . contd.

Solution:

1. **Density function for $e[k]$:** $e[k] \sim \text{GWN}(0, \sigma_e^2)$
2. **Likelihood function for $y[k]$:** We have two unknowns to estimate, $\boldsymbol{\theta} = \begin{bmatrix} c & \sigma_e^2 \end{bmatrix}^T$

$$l(\boldsymbol{\theta}|\mathbf{y}_N) = f(\mathbf{y}_N|\boldsymbol{\theta}) \tag{6}$$

Given that $e[k]$ is Gaussian, $y[k]$ also follows a Gaussian distribution with

$$E(y[k]|\boldsymbol{\theta}) = c;\ \sigma_y^2 = \sigma_e^2 \quad \Longrightarrow f(y[k]) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{1}{2}\frac{(y[k]-c)^2}{\sigma_e^2}\right) \tag{7}$$

Example 1: MLE ...contd.

Therefore,

$$f(y[0], y[1], \cdots, y[N-1]|\boldsymbol{\theta}) = \prod_{k=0}^{N-1} f(y[k]|\boldsymbol{\theta})$$

$$\implies l(\boldsymbol{\theta}|\mathbf{y}_N) = \frac{1}{(2\pi\sigma_e^2)^{N/2}} \exp\left(-\frac{1}{2}\sum_{k=0}^{N-1}\frac{(y[k]-c)^2}{\sigma_e^2}\right) \quad (8)$$

3. **Optimization problem:** The objective function to be minimized is then

$$L(\boldsymbol{\theta}, \mathbf{y_N}) = -\ln l(.) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma_e^2 - \underbrace{\frac{1}{2}\sum_{k=0}^{N-1}\frac{(y[k]-c)^2}{\sigma_e^2}}_{\text{LS obj. fun.}} \quad (9)$$

Example 1: MLE ...contd.

Setting the gradients of the objective w.r.t. c and σ_e^2 to zero, we obtain:

$$\frac{\partial L}{\partial c} = 0: \quad -\sum_{k=0}^{N-1}\frac{(y[k]-c)}{\sigma_e^2} = 0, \quad \frac{\partial L}{\partial \sigma_e^2} = 0: \quad -\frac{N}{2\sigma_e^2} + \sum_{k=0}^{N-1}\frac{(y[k]-c)}{\sigma_e^4} = 0 \quad (10)$$

Solving both equations simultaneously, we obtain the ML estimates of c and σ_e^2,

$$\boxed{\hat{c}_{\text{ML}} = \frac{1}{N}\sum_{k=0}^{N-1} y[k] = \bar{y}; \;\; \hat{\sigma}^2_{e,\text{ML}} = \frac{1}{N}\sum_{k=0}^{N-1}(y[k]-\bar{y})^2} \quad (11)$$

Example 1: MLE . . . contd.

A time-series consisting of $N = 200$ observations from a GWN process of mean $\mu = 1$ and variance $\sigma = 2$ is obtained. The ML estimates of the statistical parameters are

Estimates:	$\hat{\mu}_{\text{ML}} = 1.0467(\pm 0.1287)$;	$\hat{\sigma}_{\text{ML}} = 1.8203(\pm 0.0910)$
95% C.I.s:	$\mu \in (0.7923, 1.3012)$;	$\sigma \in (1.6619, 2.0237)$

where the values in parentheses are the 1σ standard errors in the respective estimates.

- The last term of the MLE objective function can be easily recognized as the sum square (prediction) errors in the LS formulation.
- Unlike the LS problem, MLE gives rise to a set of non-linear normal equations. Fortunately, in this case, a closed-form solution exists

Remarks on Example 1

1. The ML and LS estimates of c coincide. In general, *the ML and LS estimates of linear models coincide when the observation errors are Gaussian white noise.*
2. Estimate of variance differs slightly from the general unbiased estimator. The factor of $1/N$ in place of $1/(N-1)$ in (11) makes it statistically biased, but *asymptotically unbiased*. On the same note, the ML estimate of variance is relatively more efficient than the LS estimate.
3. The variance of c and σ_e^2 estimates are given by

$$\text{var}(\hat{c}_{\text{ML}}) = \text{var}(\bar{y}) = \frac{\sigma_e^2}{N}; \qquad \text{var}(\hat{\sigma}^2_{e,\text{ML}}) = \frac{2(N-1)}{N^2}\sigma_e^4 \tag{12}$$

Thus, the estimators are mean-square *consistent*.

Remarks . . . contd.

4. ML estimator, although being biased, *achieves the Cramer-Rao bound asymptotically*. To realize this for the problem studied above, observe from (12)

$$\text{var}(\hat{\sigma}^2_{e,\text{ML}}) \approx \frac{2}{N}\sigma_e^4 \qquad \text{(for large } N\text{)} \qquad (13)$$

5. It can be shown that the ML estimate of σ_e is

$$\hat{\sigma}_{e,\text{ML}} = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1}(y[k]-\bar{y})^2} \qquad (14)$$

Remarks on MLE: . . . contd.

6. The **asymptotic distribution** of the ML estimates are

$$\sqrt{N}(\hat{c}-c_0) \sim \text{As}\mathcal{N}(0,\sigma_e^2); \qquad \sqrt{N}(\hat{\sigma}_e^2-\sigma_e^2) \sim \text{As}\mathcal{N}(0,2\sigma^4) \qquad (15)$$

Note that the **finite sample distribution** of $\hat{\sigma}_e^2$ is a χ^2 with N degrees of freedom.

Remarks on MLE: ...contd.

7. One of the advantages of ML formulation is that *heteroskedastic* errors, *i.e.*, var$(e[k]) = \sigma_k^2$ can be naturally accommodated. Assuming the knowledge of σ_k^2, the ML estimate of $\boldsymbol{\theta} = c$ is,

$$\hat{c}_{\text{MLE}} = \hat{c}_{\text{WLS}} = \frac{\sum_{k=0}^{N-1} \frac{y[k]}{\sigma_k^2}}{\sum_{k=0}^{N-1} \frac{1}{\sigma_k^2}} \tag{16}$$

Thus, *the ML formulation also encompasses the WLS problem.*

Remarks

- The variance of the innovations estimated by ML approach is theoretically lower than that by the OLS, but it is a biased estimate. However, the estimate is *asymptotically unbiased*.
- In the OLS approach, the variance is estimated *after* estimating p model parameters. Thus effectively only $N - p$ degrees of freedom are available. In contrast, the ML approach estimates the model parameters and innovations variance jointly, resulting in lower variance.
- ML estimates are consistent and asymptotically efficient under some fairly mild conditions

Computing the MLE

The non-linear optimization problem of the MLE can be solved using Newton-Raphson and Gauss-Newton methods. In addition three other algorithms are also widely used.

1. **Fisher's scoring method:** Originally proposed by Fisher, it is a variant of the Newton-Raphson method, which replaces the computationally intensive Hessian calculations of the N-R method by their expected values
2. **Polytope method:** Unlike the gradient search approach, the idea here is to use a *heuristic* direct-search method for parameter updates Nelder and Mead, 1965. This is also known as the simplex method (different from the one in linear programming).

Computing the MLE . . . contd.

3. **EM algorithm:** The *Expectation-Maximization* algorithm formalized by Dempster, Laird, and Rubin, 1977 is an iterative methodology consisting of two steps known as the *expectation* E and *maximization* M steps, typically used for setting up the complete data likelihood from incomplete data.
   A known problem with the EM algorithm is that it does not necessarily converge to the ML estimate. However, its advantages usually outweigh the drawbacks.

For technical details, see Shumway and Stoffer, 2006 and Garthwaite, Jolliffe, and Jones, 2002.

Asymptotic properties of ML estimators

Suppose that the observation vector constitutes a random sample (GWN or i.i.d.) $\mathbf{y}$ and that the true joint p.d.f. $f(\mathbf{y};\boldsymbol{\theta}_0)$ satisfies the regularity conditions. Then the ML estimator has certain attractive asymptotic properties:

1. *Consistency*: The ML estimate converges in probability to the true parameter,

$$\hat{\boldsymbol{\theta}}_{\mathsf{ML}} \xrightarrow{p} \boldsymbol{\theta}_0 \tag{17}$$

Properties of ML estimators . . . contd.

2. *Asymptotic normality:*

$$(\hat{\boldsymbol{\theta}}_{\mathsf{ML}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_0(\boldsymbol{\theta})^{-1}) \tag{18}$$

   where $\mathbf{I}_0(\boldsymbol{\theta})$ is the Fisher's information matrix (or the inverse of Cramer-Rao's lower bound) evaluated at the true point $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

3. *Asymptotic efficiency*: A corollary of the above property is that the ML estimator asymptotically achieves the Cramer-Rao's lower bound.

$$\lim_{N\to\infty} \Sigma_{\hat{\boldsymbol{\theta}}} = \mathbf{I}_0(\boldsymbol{\theta})^{-1} \tag{19}$$

Properties of ML estimators . . . contd.

4. *Asymptotically unbiased:* ML estimators can in general produce biased estimates (with the exception of a few cases). As $N \to \infty$, this bias goes to zero.
5. *Invariance*: Suppose $\phi = g(\boldsymbol{\theta})$ for some appropriate one-to-one function $g(.)$, then a remarkable property of the ML estimator is that

$$\hat{\phi}_{\text{ML}} = g(\hat{\boldsymbol{\theta}}_{\text{ML}}) \tag{20}$$

Although these properties are listed under restrictive conditions, they hold in general for dynamic systems as well (see Ljung, 1999).

Bayesian estimation

The Bayesian estimators, named in honour of Rev. Bayes (Bayes, 1763; Stigler, 1982) deviate significantly from the classical methods discussed until now in the assumption they make on the parameter.

Assumption

Parameters of interest $\boldsymbol{\theta}$ are random with some a priori knowledge $f(\boldsymbol{\theta})$.

Remarks

- Since deterministic variables are limiting cases of random variables, the ML approach is naturally contained in these methods (we shall show shortly that MLE is mathematically the equivalent of Bayesian under some conditions).
- Strictly speaking, it is not the parameter that is random. The randomness is "imparted" due to the uncertainty in our knowledge of the parameter (both prior to and post estimation).

Basic Idea

Prior to the experiment we have a (large) uncertainty in parameters, and post experiment (and data analysis), this uncertainty should (hopefully) shrink.

The starting point is therefore, the conditional p.d.f., written using Bayes' rule:

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{y}_N)f(\mathbf{y}_N) &= f(\mathbf{y}_N|\boldsymbol{\theta})f(\boldsymbol{\theta}) \\ \implies f(\boldsymbol{\theta}|\mathbf{y}_N) &= \frac{f(\mathbf{y}_N|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{y}_N)} \end{aligned} \tag{21}$$

Posterior and Prior p.d.f.s

- The quantity $f(\boldsymbol{\theta}|\mathbf{y}_N)$ is also known as the **posterior** distribution of $\boldsymbol{\theta}$ since it is being computed *after* the observations have been obtained.
- Its companion is $f(\boldsymbol{\theta})$, known as the **prior** distribution or density. This contains the user's knowledge of the parameters prior to using the data.

The role of observed data $\mathbf{y}$ is to reduce the uncertainty in the prior information or improve our knowledge of the parameter $\boldsymbol{\theta}$.

The posterior p.d.f.

Given an observation set $\mathbf{y}_N$ (*evidence*), the denominator is a fixed quantity, allowing us to write

$$\underset{\text{(Posterior)}}{f(\boldsymbol{\theta}|\mathbf{y}_N)} = C \underset{\text{(Likelihood)}}{f(\mathbf{y}_N|\boldsymbol{\theta})} \underset{\text{(Prior)}}{f(\boldsymbol{\theta})} \tag{22}$$

The constant C is usually adjusted such that $f(\boldsymbol{\theta}|\mathbf{y})$ is a legitimate p.d.f.

Once $f(\boldsymbol{\theta}|\mathbf{y})$ is obtained, the "full" information of $\boldsymbol{\theta}$ is available, which can be used in whichever way required.

Estimation using the posterior p.d.f

- Thus, we do not have a single estimate unlike with LS or MLE, but a **range of estimates characterized by the posterior**.
- The estimator directly gives the distributional characteristics through $f(\boldsymbol{\theta}|\mathbf{y})$, in contrast to the previously studied estimators where the distribution of $\hat{\boldsymbol{\theta}}$ is constructed after computing $\hat{\boldsymbol{\theta}}$ through some approximations.
- The **point estimates** can be obtained by evaluating different properties of the p.d.f. $f(\boldsymbol{\theta}|\mathbf{y})$.

Point estimates and their optimality

- Further, these point estimates are optimal in the sense that they minimize a *risk function* $\mathcal{R}(\boldsymbol{\epsilon_\theta}) = E(\mathcal{C}(\boldsymbol{\epsilon_\theta}))$, which is the averaged user-defined *cost function*.

$$\hat{\boldsymbol{\theta}}^\star = \arg\min_{\boldsymbol{\theta}} \mathcal{R}(\boldsymbol{\varepsilon}) = \arg\min_{\hat{\boldsymbol{\theta}}} E(\mathcal{C}(\boldsymbol{\varepsilon_\theta})) \tag{23}$$

where

$$\boldsymbol{\varepsilon_\theta} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \tag{24}$$

Bayesian estimate

Three popular choice of cost functions associated with three related properties of the p.d.f. are discussed below:

1. **Bayesian estimate**, $E(\boldsymbol{\theta}|\mathbf{y})$: This is the estimate that minimizes the cost function

$$C(\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}) = ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||_2^2 \tag{25}$$

In other words, it is the MMSE of $\boldsymbol{\theta}$ and also the **mean** of $f(\boldsymbol{\theta}|\mathbf{y})$.

$$\hat{\boldsymbol{\theta}}^{\star}_{\text{BA}} = E(\boldsymbol{\theta}|\mathbf{y}) = \arg \min_{\boldsymbol{\theta}} E((\boldsymbol{\theta} - \boldsymbol{\theta}_0)^2) \tag{26}$$

Bayesian estimate . . . contd.

The support comes from the classical result in prediction theory, which states that given a random variable Y, the best prediction of an unknown X is its *conditional expectation* in the minimum MSE sense.

The estimate is computed using the definition of conditional expectation (22),

$$E(\boldsymbol{\theta}|\mathbf{y}) = \int \boldsymbol{\theta} f(\boldsymbol{\theta}|\mathbf{y})\, d\theta \tag{27}$$

MAP estimate

2. **MAP estimate**: The associated cost function is

$$C(\boldsymbol{\epsilon}_{\boldsymbol{\theta}}) = \begin{cases} 0, & |\boldsymbol{\epsilon}_{\boldsymbol{\theta}}| < \delta \\ 1, & |\boldsymbol{\epsilon}_{\boldsymbol{\theta}}| > \delta \end{cases} \tag{28}$$

This leads to the *maximum a posteriori* estimate , which is essentially the **mode** of $f(\boldsymbol{\theta}|\mathbf{y})$:

$$\hat{\boldsymbol{\theta}}^{\star}_{\text{MAP}} = \arg \max_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{y}) = \arg \max_{\boldsymbol{\theta}} f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \tag{29}$$

$$\text{or } \hat{\boldsymbol{\theta}}^{\star}_{\text{MAP}} = \arg \max_{\boldsymbol{\theta}} \left(\ln f(\mathbf{y}|\boldsymbol{\theta}) + \ln f(\boldsymbol{\theta}) \right) \tag{30}$$

The estimate is also known a the *hit-or-miss* estimate.

Remarks on MAP estimate

- When all values of $\boldsymbol{\theta}$ are equally likely a priori, i.e., $f(\boldsymbol{\theta}) = \alpha, \alpha \in \mathbb{R}^{+}$, the Bayesian estimate specializes to MLE.
- On the other hand, choosing a Gaussian prior with mean $\boldsymbol{\theta} = 0$ (or a non-zero fixed point) and variance $\frac{1}{2N\alpha}\mathbf{I}_{p\times p}$ gives an MLE with **regularization** term (see note below). Owing to this fact, the MAP estimator is also known as **penalized likelihood estimator**.

Note: Regularization is an optimization scheme where the estimation algorithm penalizes the optimizer for including more parameters than necessary by adding a penalty term to the objective function. This penalty term is usually some norm of $\boldsymbol{\theta}$.

Median estimate

2. **Median estimate:** As the name suggests, it is the **median** of the posterior.

$$\hat{\boldsymbol{\theta}}^{\star}_{\text{Median}} = \text{Median}(f(\boldsymbol{\theta}|\mathbf{y})) \tag{31}$$

The associated cost function is the absolute value of the error.

$$C(\boldsymbol{\epsilon_\theta}) = |\boldsymbol{\epsilon_\theta}| \tag{32}$$

For a Gaussian posterior, the mean, median and mode coincide.

Among the three estimators, **the MMSE is the most preferred because of its quadratic objective function, however it is one of the most difficult to compute.**

Example: Bayesian estimation of mean

We revisit the problem of estimating the mean of a signal from its measurements:

$$y[k] = c + e[k], \qquad e[k] \sim \mathcal{N}(0, \sigma_e^2) \tag{33}$$

Unlike in the case of OLS and MLE, assume some prior knowledge of $\theta \equiv c$, the uncertainty in which is described by a Gaussian p.d.f.:

$$f(\theta) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(\theta - \mu_c)^2}{2\sigma_c^2}\right) \tag{34}$$

Bayesian estimation of mean . . . contd.

Then the posterior is

$$f(\theta|\mathbf{y}) = C\frac{1}{(2\pi\sigma_c^2)^{1/2}(2\pi\sigma_e^2)^{N/2}}\exp\left(-\frac{(\theta-\mu_c)^2}{2\sigma_c^2} - \frac{\sum_{k=0}^{N-1}(y[k]-\theta)^2}{2\sigma_e^2}\right) \quad (35)$$

where the constant C is adjusted such that $\int f(\theta|\mathbf{y})\,d\theta = 1$.

Bayesian estimation of mean . . . contd.

The exponent of the posterior p.d.f. can be re-written as:

$$-\frac{\theta^2}{2}\left(\frac{1}{\sigma_c^2}+\frac{N}{\sigma_e^2}\right) + 2\theta\left(\frac{\mu_c}{\sigma_c^2}+\frac{N\bar{y}}{2\sigma_e^2}\right) - \left(\frac{\mu_c^2}{2\sigma_c^2}+\frac{\sum_{k=0}^{N-1}y^2[k]}{2\sigma_e^2}\right)$$

It is possible to show that the posterior p.d.f. is Gaussian:

$$f(\theta|\mathbf{y}) \propto \mathcal{N}(\bar{\mu},\bar{\sigma}^2) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2}}\exp\left(-\frac{1}{2}\frac{(\theta-\bar{\mu})^2}{\bar{\sigma}^2}\right) \quad (36)$$

where $\bar{\mu} \triangleq \mu_{\theta|\mathbf{y}}$, $\bar{\sigma}^2 \triangleq \sigma^2_{\theta|\mathbf{y}}$ and

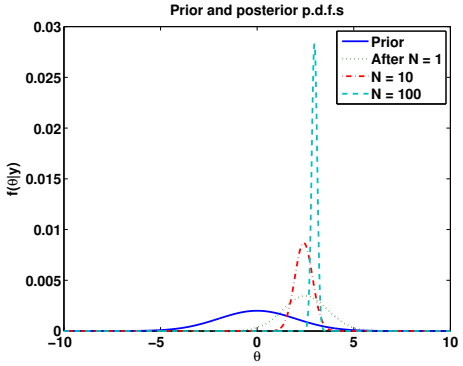
Bayesian estimation of mean . . . contd.

$$\bar{\sigma}^2 = \frac{1}{\frac{1}{\sigma_c^2} + \frac{N}{\sigma_e^2}}; \quad \bar{\mu} = \left(\frac{\mu_c}{\sigma_c^2} + \frac{N\bar{y}}{\sigma_e^2} \right) \bar{\sigma}^2 \tag{37}$$

The quantity $\bar{\mu}$ is the Bayesian estimate of c, defined in (27) and the $\bar{\sigma}$ is the standard error in this estimate.

Bayesian estimation of mean . . . contd.

Numerical illustration: Assume that $N = 100$, $\mu_c = 0, \sigma_c^2 = 4$ and that the variance of the GWN is known to be $\sigma_e^2 = 2$. The prior and the posterior p.d.f.s resulting from taking into account $N = 1, 10, 100$ observations are shown.



Observe how the data improves the highly uncertain prior knowledge to a much more precise estimate.

Remarks

1. The Bayes estimate of c is $E(c|\mathbf{y}) = \bar{\mu}$. Expressing this estimate in a weighted form provides valuable insights into the workings of the Bayesian method.

$$\bar{\mu} = \alpha\mu_c + (1-\alpha)\bar{y} \qquad \text{where} \qquad \alpha = \frac{\bar{\sigma}^2}{\sigma_c^2} = \frac{\sigma_e^2}{\sigma_e^2 + N\sigma_c^2} \tag{38}$$

   Thus, the approach gives us a weighted estimate of the prior mean (i.e., no data) and sample mean (i.e., no prior).

Remarks . . . contd.

2. The Bayesian estimate is *statistically biased*, but *asymptotically unbiased*.

$$E(\bar{\mu}) \neq \mu; \qquad \lim_{N\to\infty} E(\bar{\mu}) = \lim_{N\to\infty} \frac{\mu_c}{\sigma_c^2}\bar{\sigma}^2 + \lim_{N\to\infty} \frac{N\bar{\sigma}^2}{\sigma_e^2} E(\bar{y}) = \mu \tag{39}$$

3. The core philosophy of Bayesian estimation is that the data improves the prior estimate or knowledge of $\boldsymbol{\theta}$. This fact is established clearly in (41). Since $\beta \geq 0$ for finite N, $\bar{\sigma}^2 \leq \sigma_c^2$.

Remarks . . . contd.

4. The variance of the Bayesian estimate is a harmonic average of the variances in the estimates from two sources, namely, prior and data, respectively. Introducing

$$\beta = \frac{\sigma_c^2}{(\sigma_e^2/N)} = \frac{\text{Uncertainty in } \hat{\theta} \text{ a priori}}{\text{Uncertainty in } \hat{\theta} \text{ from data}} \tag{40}$$

we have

$$\frac{\bar{\sigma}^2}{\sigma_c^2} = \frac{1}{\beta + 1} = \alpha \tag{41}$$

Remarks . . . contd.

5. The Bayesian estimate has a lower variance than that with a non-informative prior such as MLE or LSE:

$$\bar{\sigma}^2 < \frac{\sigma_e^2}{N} \tag{42}$$

Moreover, it is a consistent estimator

$$\lim_{N \to \infty} \bar{\sigma}^2 = 0 \tag{43}$$

Closing Remarks

- Computation of the posterior can present some serious challenges depending on the choice of the (likelihood) density $f(\mathbf{y}|\boldsymbol{\theta})$ and the prior.
  - The burden is significantly reduced if the likelihood and prior form a combination such that the posterior also falls in the family of prior distributions. Then, the combination is said to form a *conjugate prior*.
  - Some well-known conjugate pairs include (Gaussian, Gaussian) (Gaussian, Gamma), (Gaussian, Wishart), (Multinomial, Dirichlet), *etc.* (Kay, 1993; Ogunnaike, 2010).

  In practice, it is important to choose the prior that is appropriate for the situation rather than going by mathematical or computational convenience.

Closing Remarks

- The core philosophy of Bayesian estimation is that the data improves the prior estimate or knowledge of $\boldsymbol{\theta}$.
- The Bayesian estimator is consistent, but a biased (asymptotically unbiased) estimator (for small samples).
  - The estimates have a lower variance than from estimators with a non-informative prior such as in MLE or LSE

Summary

- Maximum likelihood estimation technique estimates the parameters such that the likelihood function is maximized
  - The likelihood function is proportional to the p.d.f. of $\mathbf{y}$.
  - MLE is superior to and contains the LS and WLS approaches, but computationally more demanding
  - MLE gives asymptotically consistent and efficient estimates
  - The small sample performance of MLE may be inferior to that of other classical estimators.

Summary . . . contd.

- Bayesian estimators work with posterior conditional p.d.f $f(\boldsymbol{\theta}|\mathbf{y})$, which is constructed from $f(\mathbf{y})$ and the prior knowledge of parameters
  - They appeal to a larger class of problems and contain the ML estimator. Being superior to MLE, they are computationally more demanding.
  - Naturally yield interval estimates, *i.e.*, distribution characteristics of $\hat{\boldsymbol{\theta}}$
  - Point estimates can be obtained by evaluating the properties of the posterior p.d.f., popular ones being: (i) Bayes estimate (mean or the MMSE) (ii) Median and (iii) MAP (penalized MLE)

Bibliography I

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society London*, 53, p. 370.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B.* 39, pp. 1–38.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions Royal Society London A*, 222, pp. 309–368.

Garthwaite, P., I. Jolliffe, and B. Jones (2002). *Statistical Inference*. New York, USA: Oxford University Press.

Kay, S. M. (1993). *Fundamentals of statistical signal processing: Estimation theory*. Upper Saddle River, NJ, USA: Prentice Hall.

Ljung, L. (1999). *System Identification - A Theory for the User*. Upper Saddle River, NJ, USA: Prentice Hall International.

Bibliography II

Nelder, J. and R. Mead (1965). A simplex method for function minimization. *Computer Journal*, 7, pp. 308–313.

Ogunnaike, B. A. (2010). *Random Phenomena: Fundamentals of Probability and Statistics for Engineers*. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group.

Shumway, R. and D. Stoffer (2006). *Time Series Analysis and its Applications*. New York, USA: Springer-Verlag.

Stigler, S. M. (1982). Thomas Bayes' Bayesian Inference. *Journal of Royal Statistical Society, Series A*, 145, pp. 250–258.